Using collocational features to improve automated scoring of EFL texts

Yves Bestgen
Centre for English Corpus Linguistics
Université catholique de Louvain
10 Place du Cardinal Mercier
Louvain-la-Neuve, 1348, Belgium
yves.bestgen@uclouvain.be

Abstract

This study aims at determining whether collocational features automatically extracted from EFL (English as a foreign language) texts are useful for quality scoring, and allow the improvement of a competitive baseline based on, amongst other factors, bigram frequencies. The collocational features were gathered by assigning to each bigram in an EFL text eight association scores computed on the basis of a native reference corpus. The distribution of the association scores were then summarized by a few global statistical features and by a discretizing procedure. An experiment conducted on a publicly available dataset confirmed the effectiveness of these features and the benefit brought by using several discretized association scores.

1 Introduction

The importance of preformed units in language use is well established (Pawley and Syder, 1983; Schmitt, 2004; Sinclair, 1991). If some of these sequences belong to the traditional phraseological approach, signalled by their syntactic fixedness and semantic non-compositionality, the vast majority of them are conventional word combinations that display statistical idiomaticity (Baldwin and Kim, 2010; Smiskova et al., 2012). This phraseological dimension of language has important implications for learning a foreign language, as shown by many studies in applied linguistics. It not only distinguishes native speakers from non-native ones, but the number of phraseological units in a learner text is related to the overall level of proficiency in the learned language (e.g., Forsberg, 2010; Levitzky-Aviad and Laufer, 2013; Santos et al., 2012; Verspoor et al., 2012). In these studies, a limited number of expressions were analysed in a small number of texts, giving a very detailed, but also very punctual, view of the phenomenon. In addition, the phraseological nature of a lexical sequence was determined manually using dictionaries or by asking native speakers, making the analysis of numerous texts difficult.

These limitations were overcome by Durrant and Schmitt (2009), who proposed[1] assigning to the bigrams present in an EFL text two association scores (ASs), computed on the basis of a large native reference corpus: (pointwise) Mutual Information (*MI*), which favours bigrams made up of low-frequency words, and the *t*-score, which highlights those composed of high-frequency words. They observed that, compared to native speakers, EFL learners tend to underuse collocations with high *MI* scores while overusing those with high *t*-scores. More recently, Bestgen and Granger (2014, 2015) and Granger and Bestgen (2014) showed that these ASs distinguish advanced learners from intermediate learners, and that the average *MI* score and the proportion of bigrams in the text that are absent from the reference corpus were good predictors of text quality, but that the average *t*-score was much less successful. These studies have a major drawback: the effectiveness of phraseological indices was not compared to that of other features known to be effective predictors. It is therefore impossible to determine whether the phraseological indices are really effective and if they can improve the prediction when combined with other indices. This limitation is probably partly due to the fact that these analyses were not conducted in the field of automatic scoring, but in applied linguistics.

In automatic scoring, phraseological expres-

[1] See Bernardini (2007) for an earlier use of this approach in translation studies.

Proceedings of the 12th Workshop on Multiword Expressions, pages 84–90,
Berlin, Germany, August 7-12, 2016. ©2016 Association for Computational Linguistics

sions have long been used almost exclusively for detecting errors, a task for which they have been very useful (e.g., Chodorow and Leacock, 2000; Futagi et al., 2008; Wu et al., 2010). It is noteworthy that a feature tracking the correct use of collocations was considered for inclusion in e-Rater, but its usefulness for predicting text quality seems rather limited (Higgins et al., 2015). Very recently, however, Somasundaran and Chodorow (2014) and Somasundaran et al. (2015) demonstrated the benefit brought by collocation measures, amongst other linguistic features, for automatically scoring spoken picture-based narration tasks. Like Durrant and Schmitt (2009), they used a large corpus to obtain the *MI* scores of every bigram and trigram in the responses and derived a series of collocational features: the maximum, minimum and the median *MI*, and the proportion of bigrams' and trigrams' *MI* scores falling into eight bins, such as [-inf,-20],]-20,-10],]-10,-1] or]20, +inf]. They found that these features were very effective for scoring the responses, even when compared to a competitive baseline system that uses state-of-the-art speech-based features.

Even if these results were extremely promising, they leave a number of questions unanswered. First, they were obtained by studying short oral responses. Can they be generalized to longer written texts, a situation that allows the learner to spend much more time on its production? Then one can wonder whether the use of *MI* is sufficient, or if additional benefits can be obtained by taking into account other associational measures for collocations. In this context, extracting richer features than the mean scores, as done by Somasundaran and Chodorow (2014), seems particularly promising, because Granger and Bestgen (2014) found that the best learner texts contain more middle-level *t*-score bigrams and fewer low and high-level *t*-score bigrams. This observation may be related to the fact that the low *t*-score bigrams are often erroneous combinations of words, while high scores indicate extremely common bigrams in the language, which are easy to learn. It is therefore far from obvious that there is a simple linear or monotonic relationship between the distribution of the association scores (ASs) in a text and its quality. Finally, it would be interesting to determine whether using ASs extracted from a corpus of native texts enables a better prediction than that obtained by using the simple frequency of the unigrams and bigrams (Yannakoudakis et al., 2011).

This study attempts to answer these questions by extracting from the bigrams in EFL texts richer features from several association measures as described in Section 2, and by comparing the effectiveness of these collocational features to that of lexical features (Section 3). The conclusion proposes several paths for further research.

2 Extracting Collocation Features

Somasundaran and Chodorow (2014) used only one AS, while Durrant and Schmitt (2009) used two, but there are many other ASs (Pecina, 2010). Evert (2009) recommends a heuristic approach by testing a series of ASs to keep the one that is most appropriate for the task at hand, while Pecina recommends using several ASs simultaneously. These recommendations were followed here by comparing the performance of eight ASs and by combining them (i.e., using simultaneously all of them in the feature set). In addition to *MI* and *t*-score (Church et al., 1991), the six following ASs were evaluated:

1. *MI3* (Daille, 1994), a heuristic modification of *MI*, proposed to reduce its tendency to assign inflated scores to rare words that occur together,

2. *z* (Berry-Rogghe, 1973), the signed square-root of the cell contribution to the Pearson Chi-square for a 2x2 contingency table,

3. *simple-ll* (Evert, 2009), the signed cell contribution to the log-likelihood Chi-square test recommended by Dunning (1993),

4. *Fisher*'s exact test (Pedersen et al., 1996), which corresponds to the probability of observing, under the null hypothesis of independence, at least as many collocations as the number actually observed,

5. Mutual rank ratio (*mrr*, Dean, 2005), a non-parametric measure that has been successful in detecting collocation errors in EFL texts (Futagi et al., 2008),

6. *logDice* (Rychly, 2008), a logarithmic transformation of the Dice coefficient used in the Sketch Engine (Kilgarriff et al., 2014).

In order to extract more information from the distribution of the ASs in each text than the mean or the median, Durrant and Schmitt (2009) and Somasundaran et al. (2015) used a standard procedure in descriptive statistics and automatic information processing known as discretization, binning or quantization (Garcia et al., 2013). It divides a continuous variable into bins and counts the proportion of scores that fall into each bin. In their analyses, the boundaries of the bins were manually and arbitrarily defined. This approach can be used for any AS, but it makes the comparison of the effectiveness of them difficult because a weaker performance may come from a less effective AS or from poorly chosen bin boundaries. To reduce the potential impact of the choice of boundaries, a very simple and completely automatic discretization procedure was used: the Equal Frequency Discretizer, which divides the sorted values into *k* intervals so that each interval contains approximately the same number of values (Dougherty et al., 1995). It is unsupervised and depends on only one parameter (i.e., the number of bins). In the present study, it was applied separately for each AS, to every bigram present in the learners' texts and consists of two steps:

1. Partitioning the distribution of scores in bins containing the same number of bigrams,

2. Computing for each text the proportion of bigrams whose AS falls into each bin, using as a denominator the total number of bigrams in the text.

3 Experiment

To assess the benefits of relying on collocational features to predict an EFL text's quality, an experiment was conducted. This section describes the corpus used, as well as the procedures for extracting the collocational and baseline features and for scoring the texts.

3.1 Experiment Setup

Dataset: The analyses were conducted on the First Certificate in English (FCE) ESOL examination scripts described in Yannakoudakis et al. (2011, 2012). Extracted from the Cambridge Learner Corpus, this dataset consists of 1238 texts of between 200 and 400 words, to which an overall mark has been assigned. As in Yannakoudakis et al. (2011), the 1141 texts from the year 2000 were used for training, while the 97 texts from the year 2001 were used for testing.

Collocational Features: The global statistical features in Somasundaran et al. (2015) and Bestgen and Granger (2014) were used: the mean, the median, the maximum and the minimum of the ASs, and the proportion of bigrams that are present in the learner text but absent from the reference corpus. Because the best number of bins for discretizing the distributions was not known, the following ones were compared: 3, 5, 8, 10, 15, 20, 25, 33, 50, 75 and 100. To get all these features, each learner text was tokenized and POS-tagged by means of CLAWS7[2] and all bigrams were extracted. Punctuation marks and any sequence of characters that did not correspond to a word interrupt the bigram extraction. Each bigram was then looked up in the 100 million word British National Corpus (BNC[3]) and, if found, assigned its ASs. The collocational features were then computed on the basis of all the different bigrams present in each text (types) to give more weight to their diversity (Durrant and Schmitt, 2009).

Lexical Features: As a benchmark for comparison, the lexical features that were showed to be good predictors of the quality of the texts in this dataset (Yannakoudakis et al., 2011) were chosen. They consist of the frequency of the word unigrams and bigrams. This baseline is particularly relevant because it includes the lexical bigrams that are the basis of the collocational features. These features were extracted as described in Yannakoudakis et al. (2011); the only difference is that they used the RASP tagger and not the CLAWS tagger.

Supervised Learning Approach and Evaluation: As in Yannakoudakis et al. (2011), the automated scoring task was treated as a rank-preference learning problem by means of the SVM-Rank package (Joachims, 2006), which is a much faster version of the SVM-Light package used by Yannakoudakis et al. (2011). The procedure was identical to that described in their study. Since the quality ratings are distributed on a zero to 40 scale, I chose Pearson's correlation coefficient, also used by Yannakoudakis et al. (2011), as the measure of performance.

4 Results

Initial analyses focused on the interest of discretizing the ASs by assessing the benefits obtained

[2] http://ucrel.lancs.ac.uk/claws/

[3] http://www.natcorp.ox.ac.uk/corpus/

Nbin	MI	t	MI3	z	simple-ll	fisher	mrr	logDice	All
0	0.57	0.51	0.47	0.49	0.52	**0.62**	0.49	0.47	0.51
3	0.59	0.58	0.53	0.53	0.57	**0.62**	0.53	0.50	0.59
5	0.61	0.64	0.54	0.61	0.63	***0.65***	0.57	*0.51*	0.64
8	0.61	**0.63**	0.54	0.61	0.61	**0.63**	0.58	0.50	0.63
10	0.61	**0.63**	0.53	0.62	**0.63**	**0.63**	0.57	*0.51*	0.64
15	0.61	**0.64**	0.55	0.63	***0.64***	0.63	0.57	0.50	0.64
20	0.61	**0.64**	0.56	0.63	***0.64***	**0.64**	0.58	0.49	0.65
25	0.60	**0.64**	0.56	0.63	0.63	0.63	0.58	0.48	0.64
33	0.61	0.64	*0.57*	***0.65***	*0.64*	0.63	*0.59*	0.50	0.65
50	0.61	***0.65***	*0.57*	0.62	*0.64*	0.63	*0.59*	0.49	0.65
75	*0.62*	0.63	0.56	**0.64**	0.62	0.63	*0.59*	*0.51*	0.65
100	*0.62*	**0.63**	0.54	**0.63**	0.62	**0.63**	*0.59*	0.50	0.65
Mean	0.61	**0.63**	0.55	0.62	0.62	**0.63**	0.57	0.50	0.64

Table 1: Correlations for the collocational features. Note: The global statistical features are always used. The highest value on each line, ignoring the All column, is in bold type. The highest value in each column is italicized. The mean row values were computed for the different numbers of bins, disregarding the 0-bin row.

when these features were added to the global statistical features. Collocational features were then compared to the lexical features and added to them to determine the maximum level of performance that could be achieved.

4.1 Collocational Features

When no discretization procedure was used (the 0 row), *Fisher* was far more effective than the other ASs, followed by *MI*. Adding the discretized features led to far better performances (except for *logDice*), as shown by the *Mean* row. For a small number of bins, *Fisher* remained the best, but for an intermediate number, the best were *t* and *simple-ll*, and for a large number, *z* became competitive. Still, the differences between the best ASs were quite small. From eight bins and beyond, using all the ASs gave the best result, but the gain was relatively small. Regarding the number of bins, at least five seems necessary, but using many more did not harm performance. It is noteworthy that all the correlations reported in table 1 are much larger that the correlation of a baseline system based purely on length (r = 0.27).

To determine if the automatic procedure for discretizing the ASs is at least as effective as the bin boundaries manually set by Somasundaran et al. (2015), I used them instead of the automatic bins for the model with eight bins based on *MI*. The correlation obtained was 0.60, a value slightly lower than that reported in Table 1 (0.61).

4.2 Collocational and Baseline Features

The lexical features used alone allowed a 0.68 correlation[4]. These features are thus more effective than the best combinations of collocational features reported in Table 1, but, as shown in Table 2, adding the collocational features to the lexical ones produces far better performances. Steiner's t-test (Howell, 2008, p. 269-271) for comparing two non-independent correlations showed that collocational features significantly improve the prediction when compared to the baseline (all ps <0.005). If *MI* is always one of the best performing ASs, the differences between the ASs are quite low. For all numbers of bins, using all the ASs allows the best performance.

To get an idea of how well the collocational and lexical features perform, the correlations in Table 2 can be compared to the average correlation between the Examiners' scores reported by Yannakoudakis et al. (2011), which give an upper bound of 0.80 while the *All* models with more than three bins obtain a correlation of at least 0.75. Adding collocational features to lexical ones thus reduces by 58% the difference between the lexical features alone and the upper bound. However, the most difficult part of the work is still to be done.

5 Conclusion and Future Work

Following on from Durrant and Schmitt (2009), Somasundaran and Chodorow (2014) and Bestgen and Granger (2014), this study confirms the benefits conferred by collocational features for the automated scoring of EFL texts. It also shows that these features improve a competitive baseline, based among other factors on the bigram frequen-

[4] This value is higher by 0.05 than that reported by Yannakoudakis et al. (2011). As I used exactly the same procedure, the difference probably comes from the SVM-Rank/SVM-Light parameters. The SVM-Rank default settings were used except for the squared slacks for the L-norm (i.e., *-p 2*) because it provided a high performance without having to optimize other parameters such as *C*.

Nbin	**MI**	**t**	**MI3**	**z**	**simple-ll**	**fisher**	**mrr**	**logDice**	**All**
0	0.70	0.70	0.70	0.70	0.70	**0.71**	0.70	0.70	0.72
3	**0.72**	0.71	0.71	0.71	0.71	0.71	0.71	0.70	0.74
5	**0.72**	0.71	0.71	0.71	0.71	***0.72***	**0.72**	*0.71*	0.76
8	**0.72**	0.71	0.71	0.71	0.71	***0.72***	**0.72**	*0.71*	0.75
10	**0.72**	0.71	0.71	0.71	0.71	***0.72***	**0.72**	*0.71*	0.75
15	**0.72**	***0.72***	***0.72***	0.71	0.71	***0.72***	**0.72**	*0.71*	0.75
20	**0.72**	***0.72***	***0.72***	0.71	0.71	***0.72***	**0.72**	*0.71*	0.76
25	***0.73***	*0.72*	*0.72*	0.71	0.71	*0.72*	0.72	0.70	0.75
33	***0.73***	*0.72*	*0.72*	*0.72*	0.71	*0.72*	0.72	0.70	0.75
50	***0.73***	*0.72*	*0.72*	0.71	*0.72*	*0.72*	0.72	0.70	0.76
75	***0.73***	0.71	*0.72*	0.71	0.71	*0.72*	0.72	0.70	0.75
100	***0.73***	*0.72*	*0.72*	0.71	0.71	*0.72*	***0.73***	*0.71*	0.75
Mean	**0.72**	0.71	0.71	0.71	0.71	**0.72**	**0.72**	0.70	0.75

Table 2: Correlations for the collocational and lexical features. See the notes below Table 1.

cies in the texts. As proposed by Somasundaran and Chodorow (2014), binning the AS distributions improves the efficiency and, as proposed by Durrant and Schmitt (2009), considering several ASs also gives extra efficiency. Compared to Bestgen and Granger (2014), the binning allows *t* to be as effective as the *MI*. This result suggests that it might be interesting to analyse more thoroughly the complex relationship between the AS distributions in a text and its quality.

It must be kept in mind that these observations result from the analysis of a single dataset and replications are more than desirable. It is also necessary to determine whether the collocational features can improve not only the baseline used here, but also a predictive model that includes many other features known for their effectiveness. Further developments are worth mentioning. Unlike Somasundaran et al. (2015), I only used bigrams' collocational features. Whether adding trigrams would further improve the performance is an open question. Trying to answer it requires a thorough study of the association measures for n-grams longer than two words since they have received much less attention (Bestgen, 2014; Gries, 2010). It might also be interesting to evaluate other techniques to discretize the AS distributions, since this study rests on one of the simplest techniques. Further studies are also needed to better understand the impact of the combination of ASs. On the one hand, it is likely that some ASs are partially redundant and that keeping only one might be enough. On the other hand, it would be interesting to determine whether, rather than combining the AS bin proportions independently, it would be better to create the bins on the simultaneous basis of two or more ASs, such as one bin for the bigrams with high *MI* scores and medium *t*-scores.

Acknowledgments

This work was supported by the Fonds de la Recherche Scientifique - FNRS under Grant J.0025.16. The author is a Research Associate of this institution.

References

Timothy Baldwin and Su N. Kim. 2010. Multiword expressions. In Nitin Indurkhya and Fred J. Damerau, editors, *Handbook of Natural Language Processing*, pages 267–292. CRC Press.

Silvia Bernardini. 2007. Collocations in translated language. combining parallel, comparable and reference corpora. In *Proceedings of the Corpus Linguistics Conference*, pages 1–16.

Godelieve L. M. Berry-Rogghe. 1973. The computation of collocations and their relevance in lexical studies. In Adam J Aitken, Richard W. Bailey, and Neil Hamilton-Smith, editors, *The Computer and Literary Studies*. Edinburgh University Press.

Yves Bestgen and Sylviane Granger. 2014. Quantifying the development of phraseological competence in L2 English writing: An automated approach. *Journal of Second Language Writing*, 26:28–41.

Yves Bestgen and Sylviane Granger. 2015. Tracking L2 writers' phraseological development using collgrams: Evidence from a longitudinal EFL corpus. ICAME 36, Trier, May.

Yves Bestgen. 2014. Extraction automatique de collocations : Peut-on étendre le test exact de Fisher à des séquences de plus de 2 mots? In *Actes de JADT 2014*, pages 79–90.

Martin Chodorow and Claudia Leacock. 2000. An unsupervised method for detecting grammatical errors. In *Proceedings of the Conference of the North American Chapter of the Association of Computational Linguistics (NAACL)*, pages 140–147.

Kenneth Church, William A. Gale, Patrick Hanks, and Donald Hindle. 1991. Using statistics in lexical analysis. In Uri Zernik, editor, *Lexical Acquisition: Using On-line Resources to Build a Lexicon*, pages 115–164. Lawrence Erlbaum.

Paul Deane. 2005. A nonparametric method for extraction of candidate phrasal terms. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pages 605–613.

James Dougherty, Ron Kohavi, and Mehran Sahami. 1995. Supervised and unsupervised discretization of continuous features. In *Proceedings of 12th International Conference of Machine Learning (ICML)*, pages 194–202.

Ted E. Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19:61–74.

Philip Durrant and Norbert Schmitt. 2009. To what extent do native and non-native writers make use of collocations? *International Review of Applied Linguistics in Language Teaching*, 47:157–177.

Stefan Evert. 2009. Corpora and collocations. In Anke Lüdeling and Merja Kytö, editors, *Corpus Linguistics. An International Handbook*, pages 1211–1248. Mouton de Gruyter.

Fanny Forsberg. 2010. Using conventional sequences in L2 French. *International Review of Applied Linguistics in Language Teaching*, pages 25–50.

Yoko Futagi, Paul Deane, Martin Chodorow, and Joel Tetreault. 2008. A computational approach to detecting collocation errors in the writing of non-native speakers of English. *Computer Assisted Language Learning*, 21:353–367.

Salvador García, Julian Luengo, José A. Sáez, Victoria López, and Francisco Herrera. 2013. A survey of discretization techniques: Taxonomy and empirical analysis in supervised learning. *IEEE Transactions on Knowledge and Data Engineering*, 25:734–750.

Sylviane Granger and Yves Bestgen. 2014. The use of collocations by intermediate vs. advanced non-native writers: A bigram-based study. *International Review of Applied Linguistics in Language Teaching*, 52:229–252.

Stefan Th. Gries. 2010. Useful statistics for corpus linguistics. In Aquilino Sánchez and Moisés Almela, editors, *A Mosaic of Corpus Linguistics: Selected Approaches*, pages 269–291. Peter Lang, Frankfurt au Main, Germany.

Derrick Higgins, Chaitanya Ramineni, and Klaus Zechner. 2015. Learner corpora and automated scoring. In Sylviane Granger, Gaëtanelle Gilquin, and Fanny Meunier, editors, *Cambridge Handbook of Learner Corpus Research*. Cambridge University Press.

Thorsten Joachims. 2006. Training linear SVMs in linear time. In *Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD)*.

Adam Kilgarriff, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, and Vít Suchomel. 2014. The Sketch Engine: ten years on. *Lexicography*, 1(1):7–36.

Tami Levitzky-Aviad and Batia Laufer. 2013. Lexical properties in the writing of foreign language learners over eight years of study: single words and collocations. In Camilla Bardel, Christina Lindqvist, and Batia Laufer, editors, *L2 Vocabulary Acquisition, Knowledge and Use: New Perspectives on Assessment and Corpus Analysis*. Eurosla Monographs Series 2.

Andrew Pawley and Frances H. Syder. 1983. Two puzzles for linguistic theory: nativelike selection and nativelike fluency. In Jack C. Richards and Richard W. Schmidt, editors, *Language and Communication*. Longman.

Pavel Pecina. 2010. Lexical association measures and collocation extraction. *Language Resources & Evaluation*, 44:137–158.

Ted Pedersen, Mehmet Kayaalp, and Rebecca Bruce. 1996. Significant lexical relationships. In *Proceedings of the 13th National Conference on Artificial Intelligence*, pages 455–460.

Pavel Rychlý. 2008. A lexicographer-friendly association score. In *Proceedings of Recent Advances in Slavonic Natural Language Processing*, pages 6–9, Brno. Masarykova Univerzita.

Victor Santos, Marjolijn Verspoo, and John Nerbonne. 2012. Identifying important factors in essay grading using machine learning. In Dina Sagari, Salomi Papadima-Sophocleous, and Sophie Ioannou-Georgiou, editors, *International Experiences in Language Testing and Assessment—Selected Papers in Memory of Pavlos Pavlou*, pages 295–309. Peter Lang, Frankfurt au Main, Germany.

Norbert Schmitt. 2004. *Formulaic sequences: Acquisition, processing and use*. Benjamins.

John Sinclair. 1991. *Corpus, Concordance, Collocation*. Oxford University Press.

Hanna Smiskova, Marjolijn Verspoor, and Wander Lowie. 2012. Conventionalized ways of saying things (CWOSTs) and L2 development. *Dutch Journal of Applied Linguistics*, 1:125–142.

Swapna Somasundaran and Martin Chodorow. 2014. Automated measures of specific vocabulary knowledge from constructed responses (use these words to write a sentence based on this picture). In *Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–11.

Swapna Somasundaran, Chong M. Lee, Martin Chodorow, and Xinhao Wang. 2015. Automated scoring of picture-based story narration. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 42–48.

Marjolijn Verspoor, Monika S. Schmid, and Xiauyan Xu. 2012. A dynamic usage based perspective on l2 writing. *Journal of Second Language Writing*, pages 239–263.

Jian-Cheng Wu, Yuchia C. Chang, Teruko Mitamura, and Jason S. Chang. 2010. Automatic collocation suggestion in academic writing. In *Proceedings of the Association for Computational Linguistics Conference*, pages 115–119.

Helen Yannakoudakis and Ted Briscoe. 2012. Modeling coherence in ESOL learner texts. In *Proceedings of the 7th Workshop on the Innovative Use of NLP for Building Educational Applications*, pages 33–43.

Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. A new dataset and method for automatically grading ESOL texts. In *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 180–189.